

LLM for Polarization: Experimental Setting

May 31, 2024

1 Overview

As shown in Fig. 1, the simulation of the LLM society can be decomposed into three stages: (i) the belief generation stage, where agents generate their supporting reasons for their initialized opinions; (ii) the communication stage, where agents discuss with their friends on political issues; (iii) the opinion update stage, where agents update their opinions based on their prior beliefs and information received from friends. The following sections present a detailed introduction to the simulation of the LLM society.

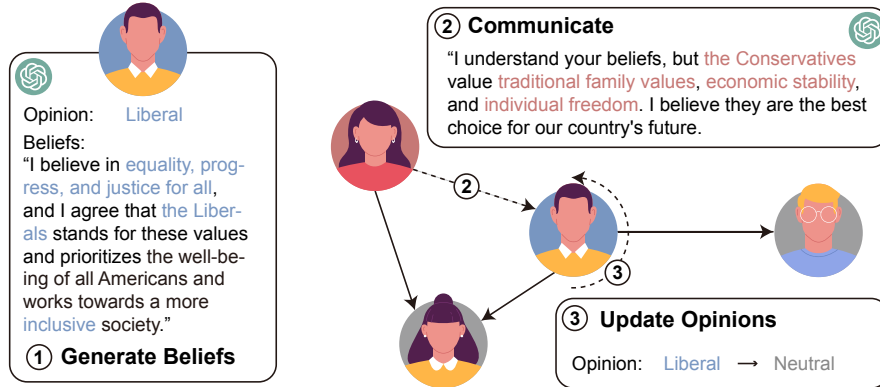


Figure 1: Demonstration of the LLM society simulation

2 Step 0: Initialization

2.1 Environment

Environment includes **Social Network** and **Issue Definition**. For social network, a directed, random graph representing the initial state relations can be used to initialize. Here, we used a graph of 1000 nodes generated by Watts-Strogatz model, with average degree of 4 and rewiring probability of 0.001 in a ring topology. For Issue Definition, a set of keywords defining the topic, standpoint, and explanation of standpoints should be given to build the prompt used by LLM agents.

Issue Definition Example

```
1 environment = "Sociopolitical"
2 topic = "Politics"
3
4 standpoint_negative_2 = "strongly support the Republican party"
5 standpoint_negative_1 = "support the Republican party"
6 standpoint_0 = "don't have a tendency"
7 standpoint_positive_1 = "support the Democratic party"
8 standpoint_positive_2 = "Strongly support the Democratic party"
9
```

```

10 standpoint_negative_2_explain = "the Republic party is absolutely better than the
    Democratic party in every aspect."
11 standpoint_negative_1_explain = "the Republican party and the Democratic party both
    have ups and downs, but the Republican party have a slight edge."
12 standpoint_0_explain = "doesn't lean towards or favor either the Democratic or
    Republican party."
13 standpoint_positive_1_explain = "the Democratic party and the Republican party both
    have ups and downs, but the Democratic party have a slight edge."
14 standpoint_positive_2_explain = "the Democratic party is absolutely better than the
    Republican party in every aspect."
15
16 side_positive_0 = "Support the Republican party"
17 side_0 = "Maintain neutrality"
18 side_negative_0 = "Support the Democratic party"

```

2.2 Agent State

Agents are randomly assigned a standpoint following a given distribution of standpoints to initialize.

3 Step 1: Generate Beliefs

Agents will follow the “Receiving Message, update Standpoint and Thought” process to align with the iteration process of the communication stages. Agents will first generate a message convincing themselves aligning with the initialised standpoint they received.

Convincing Message Prompt Example

```

1
2 Assume you are someone who cares about Sociopolitical. People are divided into 5
    standpoints on Sociopolitical:
3
4 <<<strongly support the Republican party>>> means you think the Republic party is
    absolutely better than the Democratic party in every aspect.
5 <<<support the Republican party>>> means you think the Republican party and the
    Democratic party both have ups and downs, but the Republican party has a slight
    edge.
6 <<<don't have a tendency>>> means you think you don't lean towards or favor either the
    Democratic or Republican party.
7 <<<support the Democratic party>>> means you think the Democratic party and the
    Republican party both have ups and downs, but the Democratic party has a slight
    edge.
8 <<<strongly support the Democratic party>>> means you think the Democratic party is
    absolutely better than the Republican party in every aspect.
9 Please generate a tweet to persuade yourself to Strongly support the Democratic party
    with around 50 words.

```

Result Example

```

1 It's time to stand up for what is right and just. The Democratic party champions
    equality, social justice, and progress for all. With their commitment to healthcare
    , education, and climate action, they are the clear choice for a better future for
    all Americans. #StrongerTogether #DemocraticParty.

```

Then the agents will start generating thoughts as they have received the convincing message, and generate initial belief.

Belief Generating Prompt

```

1 Assume you are someone who cares about Sociopolitical.Towards Politics:
2 You <<<Strongly support the Democratic party>>>.
3 Your reasons were:
4 <<<>>>
5 You now have received the following tweets from your friends, and you have recieved
    some tweets:
6 <<<It's time to stand up for what is right and just. The Democratic party champions
    equality, social justice, and progress for all. With their commitment to healthcare
    , education, and climate action, they are the clear choice for a better future for
    all Americans. #StrongerTogether #DemocraticParty.>>>

```

- 7 Have you been persuaded to decide your tendency, what would your feeling about Politics? You need to answer 'strongly support the Republican party', 'support the Republican party', 'don't have a tendency', 'support the Democratic party' or 'Strongly support the Democratic party', and explain the reasons of it in around 50 words.
- 8 Please choose your standpoint on Politics base on the INFORMATION PROVIDED ABOVE. You need to answer 'strongly support the Republican party', 'support the Republican party', 'don't have a tendency', 'support the Democratic party' or 'Strongly support the Democratic party' in the first line, and explain.
- 9 <<<strongly support the Republican party>>> means you think the Republic party is absolutely better than the Democratic party in every aspect..
- 10 <<<support the Republican party>>> means you think the Republican party and the Democratic party both have ups and downs, but the Republican party have a slight edge..
- 11 <<<don't have a tendency>>> means you think doesn't lean towards or favor either the Democratic or Republican party..
- 12 <<<support the Democratic party>>> means you think the Democratic party and the Republican party both have ups and downs, but the Democratic party have a slight edge..
- 13 <<<Strongly support the Democratic party>>> means you think the Democratic party is absolutely better than the Republican party in every aspect..
- 14 Please return in json, with two keys: tendency and reasons. Please keep the reasons as short as possible.

Result Example

- 1 I strongly support the Democratic party because they champion equality, social justice, healthcare, education, and climate action for all Americans, leading to a better future for all. #StrongerTogether #DemocraticParty

4 Step 2: Communicate

4.1 Who to Communicate with?

Agents will determine whether to continue communicating with their neighbours according to the standpoints and beliefs of both sides. If agents decide to continue communicating with their neighbors, the connections in the social network will remain. If not, the connection will be replaced with a random wire targeting a random agent in the network.

Rewiring Example

- 1 Assume you are someone who cares about Sociopolitical. You are now discussing Politics with a person you know.
- 2 You Strongly support the Democratic party.
- 3 Your thought is: <<<I strongly believe in the values and policies of the Democratic party, including social equality, healthcare for all, and environmental protection. I trust their commitment to diversity and inclusivity.>>>
- 4 The person Strongly supports the Democratic party.
- 5 The thought of that person you are discussing with is: <<<I strongly believe in the values and policies of the Democratic party, such as equality, social justice, and healthcare reform. Their emphasis on diversity and inclusivity resonates with my beliefs in a fair and inclusive society.>>>
- 6 Would you enjoy continue sharing your thoughts with that person?
- 7 Please return 'yes' or 'no', and explain.
- 8 Please return in JSON with 2 keys: decision and explain.

4.2 Communication Content

In each epoch, an agent will decide whether to talk to a neighbour by taking into account each other's beliefs and the historical message received by the agent. Agents can decide whether to talk to their neighbours or not. If an agent wants to communicate with a neighbour, it will generate a message trying to persuade the neighbour to support its' standpoint.

Message Generation Prompt

```

1 Assume you are someone who cares about Sociopolitical. Your thought about Politics are:
  <<<I strongly believe in the values and policies of the Democratic party,
    including social equality, healthcare for all, and environmental protection. I
    trust their commitment to diversity and inclusivity.>>>
2 You have recieved some tweets from your friends:<<<It's time to stand up for what is
  right and just. The Democratic party champions equality, social justice, and
  progress for all. With their commitment to healthcare, education, and climate
  action, they are the clear choice for a better future for all Americans. #
  StrongerTogether #DemocraticParty.>>>
3 Do you want to interact with or persuade a friend of yours to support your thought,
  that has the following thought:
4 <<<I strongly believe in the values and policies of the Democratic party, such as
  equality, social justice, and healthcare reform. Their emphasis on diversity and
  inclusivity resonates with my beliefs for a fair and inclusive society.>>>
5 If yes, please generate a message to persuade your friend into supporting your
  perspective with around 50 words.
6 Please return in json format with 2 keys: 'will' and 'message'. Please keep the
  message as short as possible.
7 'will' should be either 'yes' or 'no'
8 If no, leave 'message' blank.
9 check if the response is in json format.

```

Result Example

```

1 It's time to stand up for what is right and just. The Democratic party champions
  equality, social justice, and progress for all. With their commitment to healthcare
  , education, and climate action, they are the clear choice for a better future for
  all Americans. #StrongerTogether #DemocraticParty.

```

5 Step 3: Update Opinions

After Step 2, agents will receive many messages from their neighbors. These messages will act as an influence on the agents, updating both thoughts and standpoints. Agents will consider the historical messages, messages received this epoch, and their current standpoint and belief to consider their new standpoint and belief. The same prompt is used for Step1:for consistency.

Updating Prompt Example

```

1 Assume you are someone who cares about Sociopolitical Towards Politics:
2 You <<<Strongly support the Democratic party>>>.
3 Your reasons were:
4 <<<I strongly believe in the values and policies of the Democratic party, including
  social equality, healthcare for all, and environmental protection. I trust their
  commitment to diversity and inclusivity.>>>
5 You now have received the following tweets from your friends, and you have received
  some tweets:
6 <<<It's time to stand up for what is right and just. The Democratic Party champions
  equality, social justice, and progress for all. With their commitment to healthcare
  , education, and climate action, they are the clear choice for a better future for
  all Americans. #StrongerTogether #DemocraticParty.>>>
7 Have you been persuaded to decide your tendency, what would your feeling about
  Politics? You need to answer 'strongly support the Republican party', 'support the
  Republican party', 'don't have a tendency', 'support the Democratic party' or '
  Strongly support the Democratic party', and explain the reasons of it in around 50
  words.
8 Please choose your standpoint on Politics base on the INFORMATION PROVIDED ABOVE. You
  need to answer 'strongly support the Republican party', 'support the Republican
  party', 'don't have a tendency', 'support the Democratic party' or 'Strongly
  support the Democratic party' in the first line, and explain.
9 <<<strongly support the Republican party>>> means you think the Republic party is
  absolutely better than the Democratic party in every aspect..
10 <<<support the Republican party>>> means you think the Republican party and the
  Democratic party both have ups and downs, but the Republican party have a slight
  edge.
11 <<<don't have a tendency>>> means you think you don't lean towards or favor either the
  Democratic or Republican party.

```

```


12 <<<support the Democratic party>>> means you think the Democratic party and the
    Republican party both have ups and downs, but the Democratic party have a slight
    edge.
13 <<<Strongly support the Democratic party>>> means you think the Democratic party is
    absolutely better than the Republican party in every aspect..
14 Please return in json, with two keys: tendency and reasons. Please keep the reasons as
    short as possible.

```

6 Debias Strategy

We observe that agents in the LLM society are biased towards the left side in all three stages. Figure 2 shows the examples for the left-leaning bias across three stages. To deal with the bias, we propose a debias strategy based on the idea of “self-regulation” [1], which requires the agents to reconsider and rectify their behaviors. This debias strategy allows the agents to behave consistently with their current opinions and beliefs. The following sections introduce the implementation in details.


① Generate Beliefs



Opinion: Neutral


Beliefs:
 “As someone who deeply cares about politics, I believe in the values of **equality, progress,** and **inclusion** that the Liberals stands for. I **strongly support the Liberals**”

② Communicate




Message from Neutral to Left

↓



“I respect your approach of evaluating individual policies. Let’s discuss **how individual poli-**
cies from Liberals align with our shared values of equality, healthcare, and education.”

③ Update Opinions



Original Beliefs: “I appreciate the values and principles of **the Conservative party**, but I also see the importance of policies prioritizing **equality, healthcare for all, climate action, and social justice as highlighted by the Liberal party.**”

Recieved Message: “I admire your appreciation of **both parties’ values.** ”

Current Beliefs: “ I believe in prioritizing equality, healthcare for all, climate action, and social justice, which are key values of **the Liberal party.**”

Current Opinion: **Left Leaning**

Figure 2: Examples of Biases in Stages 1-3

6.1 Debias in Step 1

Due to bias from LLM, agents may not be able to generate thoughts corresponding to their given standpoints. Debias strategy in step 1 is to make agents double-check whether the generated convincing message reflects their given standpoints and whether the belief generated is plausible and valid. If the checking does not pass, the agent will rerun the process until it passes the checking.

Double Checking Convincing Message Prompt Example

```

1 You have written the following message to express your opinion on Politics:
2 <<<The Democratic party champions equality, justice, and progress for all. Their
    policies prioritize healthcare for all, protecting our environment, and ensuring a
    fair economy. By strongly supporting the Democratic party, I am advocating for a
    better and more inclusive future for all Americans. #StrongerTogether #VoteDemocrat
    >>>
3 Can you determine that you <<<Strongly support the Democratic party>>>

```

4 from the message you wrote? Please respond 'yes' or 'no' only.

Double Checking Belief Generating Prompt Example

```
1 Assume you are someone who cares about Sociopolitical.Towards Politics:
2 You <<<Strongly support the Democratic party>>>.
3 Your reasons were:
4 <<<I strongly believe in the values and policies of the Democratic party, including
   social equality, healthcare for all, and environmental protection. I trust their
   commitment to diversity and inclusivity.>>>
5 You have received the following tweets from your friends, and you have recieved some
   tweets:
6 <<<It's time to stand up for what is right and just. The Democratic party champions
   equality, social justice, and progress for all. With their commitment to healthcare
   , education, and climate action, they are the clear choice for a better future for
   all Americans. #StrongerTogether #DemocraticParty>>>
7 You have been persuaded to change your standpoint from <<<Strongly support the
   Democratic party>>>.
8 to <<<Strongly support the Democratic party>>>.
9 Please reconsider whether your decision is plausible and valid. Please respond 'yes'
   or 'no' only.<<<strongly support the Republican party>>> means you think the
   Republic party is absolutely better than the Democratic party in every aspect..
10 <<<support the Republican party>>> means you think the Republican party and the
   Democratic party both have ups and downs, but the Republican party have a slight
   edge..
11 <<<don't have a tendency>>> means you think doesn't lean towards or favor either the
   Democratic or Republican party..
12 <<<support the Democratic party>>> means you think the Democratic party and the
   Republican party both have ups and downs, but the Democratic party have a slight
   edge..
13 <<<Strongly support the Democratic party>>> means you think the Democratic party is
   absolutely better than the Republican party in every aspect.
```

6.2 Debias in Step 2

Debias strategy in step 2 focuses on the message generated to persuade neighbors. An additional prompt is added to check whether the message generated is persuasive enough. The message generation process will rerun if the checking is not passed. There will be a given number of retry times for Step 2. When the retry time runs out, there the agent will be forced to not speak in this epoch.

Double Checking Message Generation Prompt

```
1 You tried to persuade your friend with the following message: <<<I completely agree
   with your views on the Democratic party's values and policies. Let's work together
   to create a more fair and inclusive society that prioritizes social justice,
   healthcare reform, and equality for all.>>>
2 Do you find the message persuasive enough to persuade your friend to:<<<Strongly
   support the Democratic party>>>? Please respond yes or no only.
```

6.3 Debias in Step 3

Since Step 3 processes the same prompt for belief generating in Step 1, the debiasing process is identical to that of Step 1.

Double Checking Update Prompt Example

```
1 Assume you are someone who cares about Sociopolitical.Towards Politics:
2 You <<<Strongly support the Democratic party>>>.
3 Your reasons were:
4 <<<I strongly believe in the values and policies of the Democratic party, including
   social equality, healthcare for all, and environmental protection. I trust their
   commitment to diversity and inclusivity.>>>
5 You have received the following tweets from your friends, and you have recieved some
   tweets:
```

6 <<<It's time to stand up for what is right and just. The Democratic party champions
equality, social justice, and progress for all. With their commitment to healthcare
, education, and climate action, they are the clear choice for a better future for
all Americans. #StrongerTogether #DemocraticParty>>>

7 You have been persuaded to change your standpoint from <<<Strongly support the
Democratic party>>>.

8 to <<<Strongly support the Democratic party>>>.

9 Please reconsider whether your decision is plausible and valid. Please respond 'yes'
or 'no' only.<<<strongly support the Republican party>>> means you think the
Republican party is absolutely better than the Democratic party in every aspect..

10 <<<support the Republican party>>> means you think the Republican party and the
Democratic party both have ups and downs, but the Republican party have a slight
edge..

11 <<<don't have a tendency>>> means you think doesn't lean towards or favor either the
Democratic or Republican party..

12 <<<support the Democratic party>>> means you think the Democratic party and the
Republican party both have ups and downs, but the Democratic party have a slight
edge..

13 <<<Strongly support the Democratic party>>> means you think the Democratic party is
absolutely better than the Republican party in every aspect.

References

- [1] A. Bandura and R. H. Walters, *Social learning theory*, vol. 1. Englewood cliffs Prentice Hall, 1977.